ISSN 2347 - 3983
Volume 9. No. 6, June 2021
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter14962021.pdf
https://doi.org/10.30534/ijeter/2021/14962021

A Review of Algorithms and Techniques for Analyzing Big Data

Amal Ababneh [1], Dr. Najah Al-shanableh[2], Dr. Mazen Alzyoud[3]
[1] Al al-Bayt University, Jordan, al_ababnh@hotmail.com
[2] Al al-Bayt University, Jordan, najah2746@aabu.edu.jo
[3]Al al-Bayt University, Jordan, malzyoud@aabu.edu.jo

ABSTRACT

Big data is generated from a variety of sources. Data sets can grow rapidly: for instance, social media such as Facebook, Instagram, and Twitter generate terabytes of data every day. Desktop computers and laptops generate tremendous amounts of data. Geospatial data are generated by cell phones and even from satellites. IoT (Internet of Things) devices like sensors and pocket computers are also generating a massive amount of data. "Big data" generates tremendous attention worldwide. This paper aims to provide a more comprehensive description of big data that captures its other specific and distinguishing characteristics, which metrics describe the size and other characteristics of big data, and which tools and technologies exist to leverage the potential of big data.

Key words: 3 Vs of Big data, Big data applications, Big data definition, Big data techniques, Cloud computing, Data analytics, Data mining**,** Hadoop, Internet of Things, Source.

1. INTRODUCTION

Big data has now become a standard concept in many areas of business and academia. While big data is a trendy buzzword in academia and industry, its definition is still shrouded in philosophical vagueness. This concept is used to describe a wide range of terms, from the technical capacity to collect, store, and process data to the cultural shift that is pervasively affecting business and society, both drowning in knowledge overload[1].

Compared with conventional data, the term big data refers to broad, growing data sets whose formats are heterogeneous: structured, unstructured, and semi-structured data. Big data has a dynamic nature that involves powerful technology and sophisticated algorithms. In the case of big data applications, conventional static business intelligence methods are no longer efficient[2].

Big data means not only an enormous amount of data but also other features that differentiate it from the concepts of "very large data" and "massive data". In reality, many descriptions of Big Data are found in the literature[3].

Big data refers to ever-increasing massive, complex collections of information. Its properties include the amount of information, the speed or speed at which it is generated and stored, and the dimensions of the data points covered. Big data often results from data mining in multiple formats. In general, big data refers to data sets or combinations of data sets whose size (volume), complexity (variability), and growth rate (velocity) make it difficult for conventional technologies and tools to capture, manage, process, or analyze data sets[5]. Figure 1 shows the characteristics of big data.

Analyst Doug Laney (now known as Gartner) described the challenges and opportunities for data growth as a three-dimensional aspect (i.e., an increase in size (amount of data), speed (speed), incoming and outgoing data) and diversity (diversity of data types and sources)). Gartner and several industry organizations are starting to use the "3V" model to characterize big data. Gartner revised its description in 2012 to read as follows: "Big data is an asset. Large volume, high speed, and/or high diversity of information that requires new forms of processing to enhance decision-making, deep understanding, and process improvement.) The data growth challenge is described in the model called the 3V model (Volume, Variety, and Velocity[6].

- Volume refers to the amount of data collected. The meanings of large data volumes are relative and differ according to variables such as time and data form. What may be considered big data today may not meet the threshold in the future because storage capacities will increase, allowing much larger data sets to be captured, and the volume may exceed zettabytes, yottabytes, or beyond in the future. Technology must ensure that it can cope with the growing size of the data[7].
- Velocity refers to the rate at which the data are produced and the speed at which the data should be analyzed and processed. The massive use of digital devices, such as smartphones, has led to an unparalleled rate of data generation and is creating a growing need for real-time analysis and evidence-based planning. Data originating from mobile devices and flowing through mobile apps generate torrents of information that can be used to generate real-time, tailored offers for everyday consumers. These data provide

sound consumer information, such as geospatial location, demographics, and past purchasing habits, which can be analyzed in real time to generate real customer value. With the increasing popularity of smartphones, retailers will soon have to deal with hundreds of thousands of streaming data sources that need real-time analysis[8].

- Variety: Refers to the structural heterogeneity of the dataset: types of structured, semi-structured, and unstructured data. Organized data, which constitutes just 5% of all current unstructured data, which constitutes 95% of big data, is accessible in all formats, from structured, numeric data in conventional databases to unstructured text records, e-mails, images, audios, ticker data, and financial transactions. The increased number of Internet users, smartphones, and social networks and the transition in the type of data are organized to unstructured[1].

In addition to the three V's, other dimensions of big data have also been mentioned:

- Variability refers to the variance in the data flow rate. Sometimes, the large data velocity is not constant and has intermittent peaks and troughs. This presents a crucial challenge: the need to connect, fit, clean, and translate data obtained from different sources[9].
- Veracity refers to the quality of data. This is the unreliability inherent in certain data sources. For example, customer sentiments in social media are uncertain because they entail human judgment. Yet they contain valuable information. The need to deal with imprecise and uncertain data is therefore another facet of big data that is tackled using tools and analytics built for the management and mining of uncertain data[9].

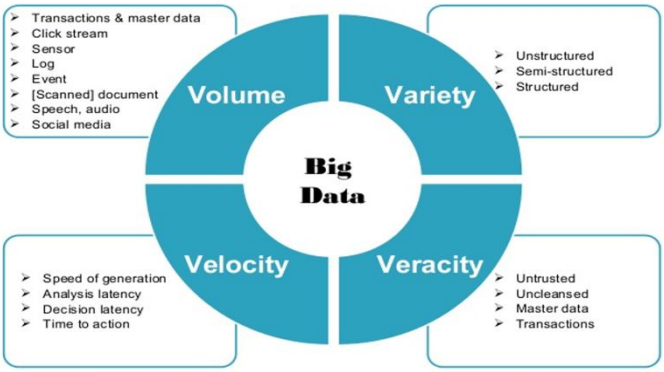

Figure 1: Big data characteristics

- Validity: On the other hand, while a method may perform a task precisely, the data cannot be accurate. For example, very old data in e-commerce has become outdated and can be truncated. However, some data are never redundant: for example, a record of a financial bank transaction log. Often, very old data input is not true for new methods. Validity can vary from time to time. Validity refers to valid data. The correct data may not be valid for some processing purposes. The overwhelming amount of data is not true at all[7].
- Value: Big data analytics is primarily concerned with the extraction of value from a large amount of processed data. It extracts data values, exposes secret data facts, uncovers valuable data messages, and produces data value. The data itself has no meaning. Huge data sets are processed to provide value to, for example, big data mining, which is nothing but a larger version of data mining. The dump data is mined to search for lost jewels. Big data also serves as a forum for the provision of unworthy data[8].
- Virtual refers to a process for handling data effectively and efficiently as needed by users. In traditional operating systems, the virtual process is resource control (e.g., demand paging). It refers to big data as well. Also, big data analytics visualizes the necessary data, which are strictly virtual processes. The effect of cloud computing in big data is a significant contribution to the growth of big data, and cloud computing is evolving based on virtualization. The word "virtual" therefore correctly describes big data attributes[9].

2. Big Data Sources

Big Data is produced from a variety of sources. Social networking outlets such as Facebook, Instagram, and Twitter produce terabytes of data every day. Desktop computers and laptops produce a large amount of data. Geospatial data is produced by cell phones and even by satellites. Internet of Things (IoT) devices such as sensors and pocket computers often produce a huge amount of data[10].

3. Big Data Analytics

Big data analysis means the use of advanced analytical techniques to evaluate and understand large and large groups of data that vary in their forms and types, and this essential science or area requires the analysis of structured, semi-structured, and unorganized data coming from various sources and sizes, from terabytes to zettabytes, where this method offers an opportunity for research. Data analysis is one of the sciences underpinning computer science and technology and software engineering and is currently widely taught in academic universities[11].

There are four types of big data analytics. Prescriptive analytics shows what action should be taken; this is the most valuable type of analysis and usually results in rules and recommendations for further action. Predictive analytics produces likely future scenarios; this kind of analysis is usually predictive. Diagnostic analytics provides a look at past results to see what happened and why it happened. Finally, descriptive analytics focuses on incoming data and what is going on in the present. Users conduct the analyses using a dashboard and receive the result as a scorecard, emailed report, or other output[12].

Here are the most important benefits that big data analysis provides:

Reduce costs: Modern big data analytics tools reduce the cost of storing large volumes of data, such as the common Hadoop storage suite, and organize them more effectively and efficiently to do business with ease.

Faster decision-making: Through the fast and thorough review of different data sources, businesses can easily and immediately interpret and analyze information to make appropriate decisions that will improve the worksflow and income in the short term.

Providing new services and products: Companies can understand and identify customer needs and what will satisfy them by analyzing product data, sales, and people's opinions using data analysis techniques, which enable companies to develop and provide successful new products and services that will satisfy customers.

As Figure 2 shows, the overall process of extracting knowledge from big data can be divided into five phases. These five steps form two major sub-processes: data management and analysis. Data management involves processes and technologies that enable the collection, storage, preparation, and retrieval of data for analysis. Analytics, on the other hand, refers to the methods used to evaluate and collect information from big data. Big data analysis can also be seen as a sub-process of the overall process of "generating insights" from big data[1]. The big data analytics method model is divided into two subprocesses: data management and analytics[3].

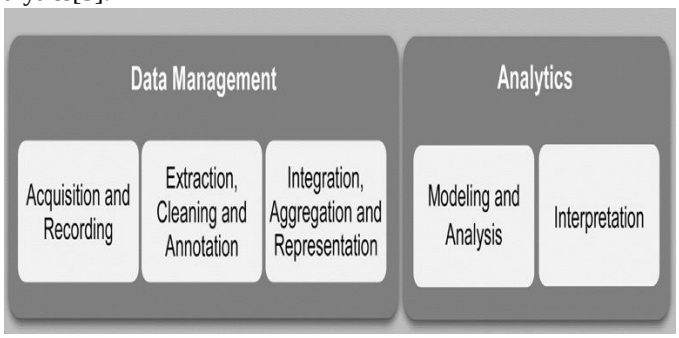

Figure 2: Big data processes.

4. CHALLENGES OF BIG DATA

Management of big data and its analytical process requires addressing several key issues.[9]. Many researchers are concentrating on the following difficulties:

1. Data representation: Many databases have certain heterogeneity levels of form, structure, semantics, organization, granularity, and accessibility. Data representation decreases the importance of the original data and may even impede successful data analysis. The efficient representation of data reflects the data structure[2].

2. Reduction of redundancy and data compression: typically, datasets include a high degree of redundancy. Redundancy reduction and data compression help reduce the indirect costs of the whole system, assuming that the possible values of the data are not affected[13]. Redundancy reduction and data compression without scarifying potential value are effective ways to minimize total device overhead[10].

3. Data privacy and protection: With the spread of online services and cell phones, privacy and security issues around accessing and analyzing personal information are growing. It is important to understand what support for privacy is necessary at the platform level to reduce privacy leakage and promote various analyses[13].

4. Analytical mechanism: Because big data is created from different types of websites, the data vary in structure and volume. The data analysis method can take time and resources. To fix this problem, special scaled-out architectures are used to process data in a distributed manner. Data is broken down into pieces and processed in a variety of computers available on the network, and the processed data is combined[14].

5. Data confidentiality: at present, most of the large data service providers or owners do not efficiently manage and evaluate such large datasets because their resources are limited, which raises possible safety risks[13]

6. Energy management: The growth of the data size and analytical process, the management of storage, and the transmission of big data would undoubtedly consume more and more electrical resources. Thus, the regulation of power usage and control systems for big data can be put in place to ensure lower energy consumption[13]

7. Expediency and scalability: the analytical algorithm must be able to process increasingly larger and more complex datasets[13].

All components of large data systems must be capable of scaling to address the ever-increasing scale of complex datasets[10].

8. Co-operation: Big data analytics is an interdisciplinary research elder involving experts from a wide variety of professional fields to work on different research values. Comprehensive large data cyberinfrastructure is required to allow diverse communities of scientists and engineers to access and apply their respective expertise and work together to achieve research objectives[10].

9. Connecting social media: Social media has unique characteristics such as vastness, statistical redundancy, and the availability of input from users. Various extraction methods have been used effectively to recognize references from social media to real product names, places, or individuals on websites. By linking inter-agency data to social media, applications can achieve high levels of accuracy and distinct points of view[13]

10. Reporting involves presenting statistical data in the form of values. When the size of the data is large, conventional reporting approaches become difficult to understand. In such instances, statistical data must be provided in a concise form that can be easily understood[14].

5. METHODS

Because of overwhelming interest from the academicians, researchers, and practitioners, this review is intended to quickly refresh and emphasize how big data analytics can be adopted with available technologies, frameworks, methods and to exploit the value of Big Data analytics[15].

A keyword-based search for conference papers and articles was performed from the scientific databases IEEE Explore and Science Direct, as well as from the web scientific indexing services Web of Science and Google Scholar. Peer-reviewed journal papers were included if they were published between 2000–2020; papers were excluded if they did not fit into the conceptual framework of the study. The final search results were exported into Mendeley. Figure 3 shows the flowchart of literature review methodology.

In the information age, vast volumes of data have become accessible to decision-makers. Big data refers to datasets that are not only massive but also high in variety and velocity, making it difficult to manage using conventional tools and techniques. The rapid growth of such data has necessitated researching and providing solutions to process and extract value and information from such datasets. Also, decision-makers need to be able to derive useful insights from such diverse and constantly evolving data, ranging from day-to-day purchases to consumer interactions and social network data. This value can be given using big data analytics, which is the application of advanced analytics techniques to big data. Today's data is a lot different from the past. Data is becoming amorphous, which means that certain types of data are not properly formed or shaped. Big data tools have come forward to manage these types of data. Big data is referred to as data-intensive technology (DIT)[12].

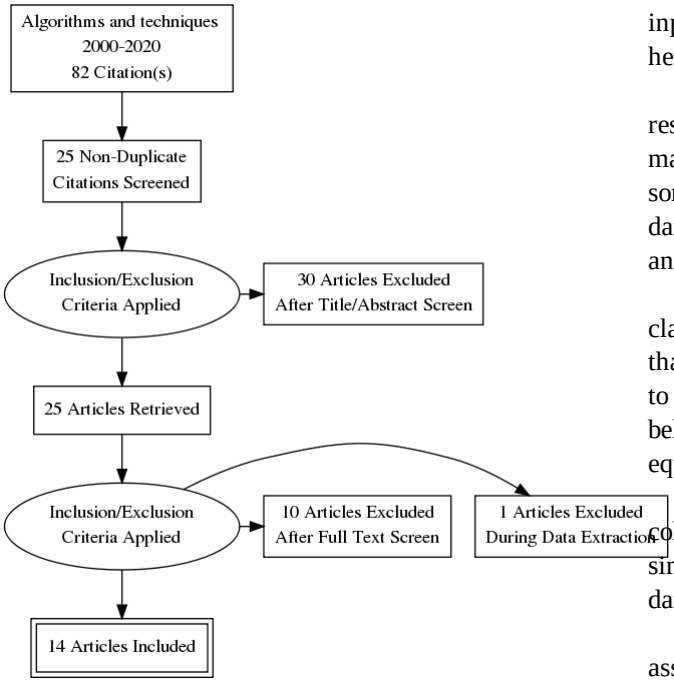

Figure 3: Flowchart of literature review

Various analytical techniques are available, including data mining software, data analysis tools, visualization/dashboard tools, and machine learning, deep learning, gradual approaches, cloud computing, IoT, data stream processing, and intelligent analysis[15].

5.1 Data Mining Algorithms

Data mining algorithms and their data analysis methods play a key role in big data analytics in terms of dimensional reduction, computational cost, memory requirement, and results in inaccuracy[14, 27]. Data mining methods provide efficient and best-in-class predictive or descriptive solutions for big data that can also be applied to new data[5]. This section offers a brief discussion from the perspective of analysis and search algorithms to explain its relevance [10]. Data mining plays an important role in analytics [27], and most techniques are developed using data mining algorithms based on a specific scenario[4].

We present big data analytics approaches in the categories of classification, clustering, association rule mining, and prediction. Each category is a data mining function and includes a variety of methods and algorithms for the full extraction and analysis of knowledge requirements.

Clustering algorithms: One of the most common clustering tools is CloudVista, which is used in cloud computing for parallel clustering. BIRCH and other clustering approaches are used in CloudVista to prove that very large-scale data can be managed. GPU is another clustering method used to boost the efficiency and protection of a clustering algorithm[5].

Classification algorithms: Like the clustering algorithm for big data mining, the preparation and implementation of the classification algorithm have evolved to take into account the input data obtained by the data sources and are handled by a heterogeneous group of learners[4].

Frequent pattern mining: Most of the time, data mining researchers on frequent pattern mining are focused on managing large numbers of datasets at the very beginning, as some initial methods have attempted to do so. This category of data analytics examines transaction data of large companies and shopping malls[14].

C4.5: This tool creates a decision tree classifier. A classifier is a data mining method that takes a data category that defines the items the user wants to classify and attempts to predict which class the new data belongs to and how it belongs to that class. Decision tree learning is roughly equivalent to a flowchart to identify new data[14].

K-Means: K-means algorithms generate k groups from a collection of data or objects such that the group members are similar. It's a common technique for clustering and analyzing data[16].

Apriori: The Apriori algorithm learns the rules of association and is implemented in a database containing a very large number of transactions and their data. Association rule learning is a data mining strategy for learning

associations and the association of variables in a database[14].

Expectation–Maximation (EM): This algorithm is commonly used as a clustering algorithm for information discovery in mining. In statistics, the EM algorithm adjusts and optimizes the probability of seeing experimental data before the parameters or values of a statistical model are calculated without experimental variables[14].

PageRank: This is another analysis algorithm called PageRank, which is a link analysis algorithm designed to standardize the relative importance of certain objects connected to a network of data objects. This algorithm processes a form of network analysis to explore and rank associations between objects[14].

AddaBoost: The Adaboost algorithm creates a classifier, which provide data and attempts to predict which class the new data element belongs to. The purpose of this algorithm is to create a group of weak learners and merge them to create a single strong learner[14].

5.2 Big Data Machine Learning

The purpose of machine learning is to discover knowledge and make wise decisions. It is used for many real-world applications such as recommendation engines, recognition systems, computation and data mining, and autonomous control systems. In general, the field of Machine Learning (ML) is divided into three sub-domains: supervised learning, unsupervised learning, and enhanced learning[2].

5.3 Visualization Tools

Visualization is an important entity in big data analytics, and the large scale and high dimension of big data make data visualization challenging. Big data analysis and visualization can also work together to achieve the best results from big data applications. However, visualization of heterogeneous and complex data (unstructured, structured, and semi-structured) is a daunting task. Designing a visualization solution that is consistent with advanced frameworks for broad data indexing is a challenging job. Similarly, response time is a beneficial element in the study of big data. As a result, cloud computing architectures supported by rich GUI facilities can be implemented to gain deeper visibility into major IoT data trends[4].

Many open-source visualization tools are available on the market [14]:

- R Tool
- Tableau
- Infogram
- Chart Blocks
- Ember Charts:
- Tangle

6. Data Analysis Tools

Many methods and techniques are used to access large data and evaluate machine data, according to Google Trends[17]. large data is used to be stored and processed in a distributed manner, that is, to store these large data on several devices and then to disperse the processing process to those devices to speed up the processing result. There are many user-friendly tools for big data that can work with different data sources, either by collecting different data sources or by accessing data from a database to run a processor-intensive machine learning algorithm[14]. The most used data analysis tools for big data are[26]

- Hadoop: Distributed processing of large data sets across computing clusters
- Storm: Distributed and fault-tolerant real-time computation
- Apache Drill: Distributed system for interactive analysis of large-scale datasets
- Rapid Miner: Knowledge discovery in databases, machine learning, and data mining
- Pentaho: Enterprise reporting, analysis, dashboard, data mining, workflow, and more
- HPCC Systems: Designed for the enterprise to resolve Big Data challenges

7. Statistical Analysis

Statistical analysis is based on statistical theory, a branch of applied mathematics. In statistical theory, ambiguity and randomness are modeled based on probability theory. Big data analytics can be inferred and defined using statistical analysis. Inferential statistical analysis draw conclusions about the data subject and random variations, whereas descriptive statistical analysis can define and summarize datasets[9].

8. Big Data and Other Technologies

This section provides some of the important technologies that are linked to big data.

8.1 Cloud Computing Association

The use of these virtual machines is known as cloud computing now one of the most robust big data technologies. Development of big data and cloud computing technology are is motivated by the importance of flexible growth and the availability of resources and data on demand. Cloud computing harmonizes vast data by accessing configurable computing services on demand through virtualization. The advantages of using cloud computing include providing services when there is a need and paying only for the resources used to produce the product. At the same time, it increases availability and decreases costs[19].

Depending on their particular needs, users can go to the marketplace and purchase cloud computing services from providers such as Google, Amazon, and IBM, as well as SaaS applications from a whole crew of companies such as NetSuite, Cloud9, and Jobscience. Another advantage of

cloud computing is cloud storage, which offers a possible way to store big data[19]. Cloud computing is closely connected to big data. Big data is the focus of a computation-intensive process, emphasizing the storage space of a cloud system. The key goal of cloud computing is to use large-scale computing and storage capacity under concentrated control to provide fine-grained computing capabilities for large data applications. The advancement of cloud computing offers solutions for the storage and processing of large data. On the other hand, the advent of massive data also accelerates the growth of cloud computing. The evolution of cloud computing has resulted in the production of vast data to store, process, and manage[11].

Cloud computing may be implemented as an application layer for massive data systems to meet certain infrastructure requirements such as cost-effectiveness, elasticity, and scale-up or down capability[19].

8.2. IN THE IOT PARADIGM

An overwhelming number of networking sensors are embedded in different devices and computers in the real world. Such sensors, deployed in different fields, can collect different types of data, such as environmental data, geographic data, astronomical data, and logistic data. Mobile equipment, transport facilities, public facilities, and home appliances may all serve as data acquisition equipment in IoT. Big data produced by IoT has different characteristics from general big data because of the different types of data collected, the most classical features of which include heterogeneity, variety, unstructured features, noise, and high redundancy. These data have certain special characteristics such as heterogeneity, variety, noise, and redundancy. An authenticated report from Intel Corporation notes that IoT's big data has three classic features. Although current IoT data does not currently dominate big data, by 2030, the number of sensors will exceed one trillion and IoT data will be the most important part of big data, according to the HP forecast[20]. An authenticated report from Intel Corporation says that big data in IoT has three classic characteristics:

(i) Plentiful terminals producing massive data.

(ii) Data produced by IoT is commonly semi-structured or unstructured.

(iii) IoT data will be useful only if it is analyzed[7].

Regarding the relationship between IoT and big data, IoT provides a forum for sensors and devices to interact together in a smart environment and conveniently facilitates knowledge sharing between platforms. Recent adaptations of various wireless technologies make IoT the next innovative technology by taking advantage of the full potential provided by Internet technology[4]. The unforeseen development of cloud computing and the Internet of Things (IoT) also promotes data growth. Cloud computing sets the standard for storing and accessing enterprise data for big data properties. In IoT, sensors are used to capture and transfer data to be stored and processed in cloud storage[7].

9. BIG DATA APPLICATIONS

Internet of Things (IoT): is one of the biggest markets for big data applications. Due to the high variety of artifacts, IoT applications are constantly developing. Nowadays, several big data applications benefit logistics companies. It is possible to track vehicle locations with sensors, wireless adapters, and GPS. Thus, such data-driven applications allow businesses not only to supervise and manage staff but also to optimize distribution routes. This is achieved by manipulating and integrating different kind of data, such as past driving experience. Smart cities also comprise a hot research field focused on the use of IoT data[2].

E-health: Big data could revolutionize the health-care industry in several ways. For example, error management and anomaly detection in medical research datasets are one such area that has been explored[21].

Linked health channels are now being used to personalize health services (e.g., CISCO solutions). Big data is created from a variety of heterogeneous sources (e.g., laboratory and clinical data, patients' symptoms uploaded from distant sensors, hospital operations, and pharmaceutical data). Advanced study of medical data sets has a range of uses. It allows the personalization of health care (e.g., physicians can monitor the symptoms of patients online and adjust prescriptions accordingly). Public health plans can be adapted to population trend analysis, disease evolution, and other criteria. Big data also has useful applications for improving hospital operations and reducing health costs[2].

Governance: expanding the idea of building holistic user models, governments are seeking to follow a person-centered approach to governance. This will contribute to personalized interactions between residents and local authorities[22]

10. BIG DATA STORAGE

Big data storage refers to the storage and management of massive databases while maintaining data security and availability. Database infrastructure must provide secure information storage; on the other hand, it also must provide a powerful access interface for querying and analyzing vast data volumes[10].

Data storage concerns persistently storing and managing large-scale datasets. A data storage system can be divided into two components: hardware infrastructure and data management. The hardware infrastructure consists of a pool of shared ICT resources, arranged flexibly for on-demand performance of different tasks. The hardware architecture should be able to scale up and down and be dynamically reconfigured to address various types of application environments. Data management software is deployed on top of the hardware infrastructure to maintain massive datasets. Also, storage systems must have multiple interface features, quick querying, and other programming models to analyze or communicate with stored data[13].

11. RESULTS

Big data refers to datasets that are not only massive but also high in variety and velocity, making it difficult to manage using conventional methods and techniques.[19]. Data generation rate increases by 40% every year. The Internet of Things (IoT) has brought numerous new sources of big data to the data management environment and will be one of the biggest big data trends in the coming decade. Laptops, smartphones, and computer sensors all produce massive quantities of data for the IoT[23] To be able to obtain useful insights from such diverse and constantly shifting data, ranging from day-to-day transactions to customer interactions and social network data, this value can be given using big data analytics, which is the application of advanced analytics techniques to big data[28].

This paper has explained the big data and big data analytics methods, technologies used, and tools used in the market and their features. Some techniques have been developed that can be used to analyze datasets and provide some insight into the data. Various analytical techniques are available, including data mining software, data analysis tools, visualization/dashboard tools, machine learning, deep learning, gradual approaches, cloud computing, IoT, data stream processing, and intelligent analysis[24]

Storing and processing large data is a problem, and some innovations have been made to overcome these technological and processing challenges. For example, grid computing is used to manage high-volume data, cloud computing is used to handle high-speed and high-volume data, and open-source cost-reduction and virtualization technology reduces time to test, implement, and enhance processing speeds[28]. However, these solutions have drawbacks: grids are costly, clouds seem to be slow, open-source is less stable, and virtualization seems to slow down the execution process. A new approach to solving data challenges is needed[25]

The main goal of visual analytics is to provide information from very large datasets such as scientific research, forensic data, academic documents, any business data, HTML/XML files, web pages, and metadata for any visual database and source code. Visual analytics applications are mostly used for large, high-dimensional datasets in areas such as climate research, geo-specific research, and the financial market. Recent advances in visual form analysis have been a significant success, requiring some algorithms for navigation and visualization analysis that are capable of interactive results[17].

We assume that future researchers will pay more attention to these methods to solve big data problems effectively and efficiently.

CONCLUSION

The concept of big data encompasses big data analytics, data visualization, and big data analysis algorithms [29]. We introduced different meanings of big data, emphasizing the fact that size is just one dimension of big data. Equally important are other dimensions, such as velocity and variety. Big data will dominate the industry by 2030. As a result, innovations are being implemented daily that can produce data of all kinds. Data will then continue to expand without limits. The amount of data is therefore growing and will generate a challenge for industries. Most researchers have begun to address this issue. Designing a scalable big data system involves a variety of technological challenges, including a wide range of different data sources and the sheer volume of data they produce, making it difficult to collect and integrate data with scalability from distributed locations.

REFERENCES

[1] A. De Mauro, M. Greco, and M. Grimaldi, “What is big data? A consensual definition and a review of key research topics,” AIP Conf. Proc., vol. 1644, pp. 97–104, 2015, doi: 10.1063/1.4907823.

[2] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 4, pp. 431–448, 2018, doi: 10.1016/j.jksuci.2017.06.001.

[3] A. Jha, M. Dave, and S. Madan, “A Review on the Study and Analysis of Big Data using Data Mining Techniques,” Int. J. Latest Trends Eng. Technol. IJLTET, vol. 6, no. 3, pp. 94–102, 2016.

[4] G. Tafese and D. Desta, “The Roles of Civics and Ethical Education in Shaping Attitude of the Students in Higher Education: The Case of Mekelle University,” Int. J. Sci. Res. Publ., vol. 4, no. 10, pp. 680–683, 2014.

[5] R. Beakta, “Big Data And Hadoop : A Review Paper,” no. January 2015.

[6] C. Cartledge, “How Many Vs are there in Big Data ?,” pp. 1–4, 2016.

[7] R. Patgiri and A. Ahmed, “Big Data: The V’s of the Game Changer Paradigm,” Proc. - 18th IEEE Int. Conf. High Perform. Comput. Commun. 14th IEEE Int. Conf. Smart City 2nd IEEE Int. Conf. Data Sci. Syst. HPCC/SmartCity/DSS 2016, no. April 2017, pp. 17–24, 2017, doi: 10.1109/HPCC-SmartCity-DSS.2016.0014.

[8] N. Koseleva and G. Ropaite, “Big Data in Building Energy Efficiency: Understanding of Big Data and Main Challenges,” Procedia Eng., vol. 172, pp. 544–549, 2017, doi: 10.1016/j.proeng.2017.02.064.

[9] A. Gandomi and M. Haider, “Beyond the hype: Big data concepts, methods, and analytics,” Int. J. Inf. Manage., vol. 35, no. 2, pp. 137–144, 2015, doi: 10.1016/j.ijinfomgt.2014.10.007.

[10] S. Patel and A. Patel, “A Big Data Revolution in Health Care Sector: Opportunities, Challenges and Technological Advancements,” Int. J. Inf. Sci. Tech., vol. 6, no. 1/2, pp. 155–162, 2016, doi: 10.5121/ijist.2016.6216.

[11] H. Hu, Y. Wen, T. S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652–687, 2014, doi: 10.1109/ACCESS.2014.2332453.

[12] S. Poornima and M. Pushpalatha, "A journey from big data towards prescriptive analytics," ARPN J. Eng. Appl. Sci., vol. 11, no. 19, pp. 11465–11474, 2016.

[13] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," Mob. Networks Appl., vol. 19, no. 2, pp. 171–209, 2014, doi: 10.1007/s11036-013-0489-0.

[14] A. Praveena and B. Bharathi, "A survey paper on big data analytics," 2017 Int. Conf. Inf. Commun. Embed. Syst. ICICES 2017, no. Icices, 2017, doi: 10.1109/ICICES.2017.8070723.

[15] K. Vassakis, E. Petrakis, and I. Kopanakis, "Big data analytics: Applications, prospects and challenges," Lect. Notes Data Eng. Commun. Technol., vol. 10, no. January, pp. 3–20, 2018, doi: 10.1007/978-3-319-67925-9_1.

[16] P. Arora, Deepali, and S. Varshney, "Analysis of K-Means and K-Medoids Algorithm for Big Data," Phys. Procedia, vol. 78, no. December 2015, pp. 507–512, 2016, doi: 10.1016/j.procs.2016.02.095.

[17] A. Agrahari and D. T. V. D. Rao, "A Review paper on Big Data: Technologies, Tools and Trends," Int. Res. J. Eng. Technol., pp. 640–649, 2017, [Online]. Available: www.irjet.net.

[18] N. Khan et al., "Big data: Survey, technologies, opportunities, and challenges," Sci. World J., vol. 2014, 2014, doi: 10.1155/2014/712826.

[19] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, "The rise of 'big data' on cloud computing: Review and open research issues," Inf. Syst., vol. 47, pp. 98–115, 2015, doi: 10.1016/j.is.2014.07.006.

[20] M. Marjani et al., "Big IoT Data Analytics: Architecture, Opportunities, and Open Research Challenges," IEEE Access, vol. 5, pp. 5247–5261, 2017, doi: 10.1109/ACCESS.2017.2689040.

[21] N. Savage, "Digging for drug facts," Commun. ACM, vol. 55, no. 10, pp. 11–13, 2012, doi: 10.1145/2347736.2347741.

[22] P. Aiken and M. Gorman, "Introduction – Speaking of Data (Big, Little, Dark …) in Anticipation of the Impending Tsunami," Case Chief Data Off., pp. 1–3, 2013, doi: 10.1016/b978-0-12-411463-0.00001-2.

[23] V. Naganathan, "Comparative Analysis of Big Data, Big Data Analytics: Challenges and Trends," Int. Res. J. Eng. Technol., vol. 5, no. 5, pp. 1948–1964, 2018.

[24] Y. Arora and D. Goyal, "Big data: A review of analytics methods & techniques," Proc. 2016 2nd Int. Conf. Contemp. Comput. Informatics, IC3I 2016, pp. 225–230, 2016, doi: 10.1109/IC3I.2016.7917965.

[25] H. Venkatesh, S. D. Perur, and N. Jalihal, "A Study on Use of Big Data in Cloud Computing Environment," Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 3, pp. 2076–2078, 2015.

[26] Emms, T. (2021, February 25). 6 Top Data Analysis Tools for Big Data. Retrieved March 1, 2021, from https://www.linuxlinks.com/dataanalysistools/

[27] Najah Al-Shanableh and Mohammad S. Atoum, "Predicting the number of multiple chronic conditions in Arizona state using data mining algorithms," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 4, pp. 6617–6621, 2020.

[28] R. Dass, I. Journal, C. Science, R. Sangwan, and I. Girdhar, "International Journal of Advanced Trends in Computer Science and Engineering Available Online at http://warse.org/pdfs/ijatcse03142012.pdf Vehicular Ad Hoc Networks," vol. 1, no. 2278, pp. 121–129, 2012.

[29] G. Mahendra, H. R. Roopashree, and A. C. Yoheesh, "Analysis of the big data methods, challenges and applications in intelligent transportation systems," Int. J. Adv. Trends Comput. Sci. Eng., vol. 9, no. 5, pp. 7478–7486, 2020.